第 9 卷第 3 期 信息技术快报 Vol.9 No.3
Information Technology Letter May 2011

面向信息过滤的多通道网络流分类研究

王鹏 李军 张鹏

摘要：随着信息技术的飞速发展，信息安全问题越来越得到全社会的重视。其中网络内容安全是最突出的问题之一，而作为网络内容安全处理核心技术的网络数据流过滤技术也面临着新的挑战。本文从网络数据流过滤问题出发，研究利用多通道信息进行网络数据流分类的技术，包括以下三方面的工作：（1）多通道网络流分类模型研究，提出了可融合网络结构信息和网络内容信息的流分类模型；（2）分类模型索引技术研究，提出一种基于 R-Tree 分类模型索引结构，极大地提高了网络数据流的判别速度；（3）多通道网络流过滤系统 F9 实验平台建设，该系统支持多通道网络流判别过滤，可作为新模型与算法的实验平台。以上三方面的工作从模型构造，模型索引，和模型实现三方面系统研究了面向信息过滤的多通道网络流分类系统。

关键词：多通道网络流 信息过滤 数据流分类 模型索引 F9 过滤系统

1 背景与研究意义

近年来，随着信息技术的普及和发展，互联网已经深入社会生活的方方面面，随之而来的利用网络来传播反动、色情内容等恶意信息的问题也越发严重。因此对网络数据流进行过滤分类具有重大意义。网络数据流包括了多种不同结构的信息，如 IP 信息，网页地址（URL，Uniform Resources Locator）信息以及文本图片信息等多媒体内容，传统的网络内容安全处理技术往往只利用网络中单一信息进行过滤分类，如利用网页文本或网页地址对网络流进行分类，这种方法准确度差，并且极易被破解（如将文本嵌入到图片中），难以达到实际应用的需求。而利用网络数据流多通道信息进行分类过滤，因其精确度高并且难以破解的优势，已经得到学术界和工业界的广泛关注，并成为网络安全领域的研究热点。

多通道网络流是指在网络访问中，一个网络请求所对应的网络内容信息（比如文本流、图片流、视频和音频流等）和网络结构信息（比如网页地址链接、IP 地址、协议类型等）的总和。由于用户的网络访问行为由多通道信息构成，为了判别一个用户访问是否包含非法信息，可以联合这些多通道信息进行综合判定。

多通道网络流
页面请求
请求的网页
网页所在的
服务器
文本
图片
视频
音频
IP地址
域名服务器
协议类型
网址链接

图 1. 多通道网络流示意图

信息过滤的核心问题是如何构造多通道网络流上的精确分类模型，以对未来任意未知的流量进行准确判别。其中，一种有效的方法是基于结果融合的多通道网络流分类模型，该类模型首先提取各个通道上的特征信息，然后在每个通道上分别构造分类器，最后把这些分类器的结果融合起来进行综合判断。

2　相关工作

面向过滤的多通道网络流分类技术是网络内容安全处理的核心技术，它与深度分组检测（Deep Packet Interception，DPI）技术[11]，字符串匹配技术[12]，信息抽取技术[13]，多媒体特征抽取技术，数据索引技术，数据库技术等紧密相关。学术界与工业界有很多工作从不同的角度来阐述和解决这个问题。

网络流过滤问题可以看作是一个数据分类问题。VFDT（Very Fast Decision Tree，快速决策树）[1]是专门针对数据流分类的决策树，随着数据流的增长，以增量方式与数据流模式相对应生成。但是 VFDT 无法处理数据流概念漂移的问题，CVFDT (Concept-adapting Very Fast Decision Tree，概念调整快速决策树)[2]解决了这个问题。它不断对决策树进行裁剪，并应对新的模式生成新的决策分枝，这很好地解决了数据流概念漂移的问题。VFDTc[3]对 VFDT 进行扩展，使其可以处理连续属性，并可应对概念漂移情况。这样的数据流分类方法，人工无法定制过滤规则，因此不适于用信息安全领域的网络流过滤。

当前，国际上也已有专门针对数据流进行管理和判别的应用系统。其代表有斯坦福大学的 STREAM（STanford stREam datA Manager）系统[4]。它支持类似 SQL[1]语言的数据流查询语言 CQL（Continues Query Language）[5]。通过 CQL 可以注册连续查询，对数据流进行查询操作。它支持 SQL 语言的大部分语法，由于数据流的特殊性，这样的查询总是针对某一时间窗口的，并且返回的结果是近似值。如果将注册的连续查询看作是分类规则的话，STREAM 系统是一个支持复杂规则的网络流分类过滤系统，但是 STREAM 系统缺少对复杂过滤器的支持。

尽管现有的解决方案众多，但仍存在以下几点不足：(1)这些解决方案只能处理单一形式的数据流，无法处理多通道网络数据流；(2)在数据流分类的研究上大都只考虑如何建立分类模型，关注点是分类精度，而不考虑分类的速度。然而，在高速网络流过滤分类问题上，分类的速度与精度一样重要；(3) 目前开发的系统都在模拟环境下进行测试，缺少在真实网络环境下的大流量和高强度测试。

3　我们的研究工作介绍

围绕多通道网络流分类的模型构造、模型索引和系统开发这三方面问题，我们开展了多通道网络流上的集成分类模型研究、基于 R-树[2]的分类模型索引技术研究、以及 F9 实验平台建设。其中，分类模型的研究主要解决网络流分类中的基础理论问题；模型索引方面的研究主要解决模型的实时判别问题；F9 系统开发工作则主要解决模型测试和应用转化问题。这些工作互相促进，构成了一个整体。

3.1 面向过滤的多通道网络流分类模型

当融合各个通道的信息的时候，每个通道上的决策可能会相互矛盾。如图 2 所示，对于一个包含多媒体内容的网页，从内容层面上来看，文本通道上的判别结果和图片通道上的结果可能不同。从网络结构的角度来看，IP 地址通道或者网页地址通道也可能给出不同的决策结果。总之，当融合各个通道的信息的时候，一个不可避免的问题就是各个通道上的决

[1] Structured Query Language，结构化查询语言

[2] R-Tree（real-tree），是 B-tree 向多维空间发展的一种形式，它将空间对象按范围划分，每个结点都对应一个区域和一个磁盘页，是目前流行的空间索引。详见本文§3.2

策矛盾。

此外，当把多个不同通道上的判别器进行融合的时候，从决策模型的角度必须考虑到：（1）各个判别器的领域不同——它们可能分别针对文本、图像、网页地址等不同对象；（2）各个判别器的判别能力不同——由于在数据流上的训练成本很高，因此，我们往往只能构造一小部分的分类器，而构造聚类器则较为容易；（3）各个判别器的判别能力会随时间变化——由于数据流是连续变化的，各个判别器对其的判别能力一般会随时间衰减。我们提出了一个聚类器和分类器组合的决策模型，从决策融合的角度采用以下三步来解决以上三个问题。



图 2. 多通道决策矛盾问题

第一步：**决策编码** 这实际上是解决不同模型之间的相似性度量问题。举个例子，假设需要把一个网页判别为三类（风险高、风险中、风险低）中的一类，我们训练了四个不同类型的判别器 λ^1、λ^2、λ^3、λ^4，其中前两个是分类器，后面两个是聚类器。对于当前到来的 7 个网页 x_1，……，x_7，假设得到的判别结果如表 1 所示。其中 g_i 是对应的编码的基。由于是三类分类问题，每个模型都使用三个基，每个样本都对应一个坐标，比如网页 x_1 被模型 λ^1 分为第一类，它的坐标为[1, 0, 0]。这样，我们就可以通过杰卡德距离[3]（Jaccard distance）来度量两个基之间的相似性。比如基 g_4 和 g_8 的相似性是 2/3，而 g_8 和 g_9 的相似性是 1/5，也就是说，g_4 和 g_8 更加相似。

	类/聚类 标识符				组向量											
					λ^1			λ^2			λ^3			λ^4		
	λ^1	λ^2	λ^3	λ^4	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}	g_{12}
x_1	1	1	2	1	1	0	0	1	0	0	0	1	0	1	0	0
x_2	1	1	2	2	1	0	0	1	0	0	0	1	0	0	1	0
x_3	1	2	1	3	1	0	0	0	1	0	1	0	0	0	0	1
x_4	2	2	3	1	0	1	0	0	1	0	0	0	1	1	0	0
x_5	3	2	3	2	0	0	1	0	1	0	0	0	1	0	1	0
x_6	3	3	1	1	0	0	1	0	0	1	1	0	0	1	0	0
x_7	2	1	3	1	0	1	0	1	0	0	0	0	1	1	0	0

表 1 分类器和聚类器的相似性度量

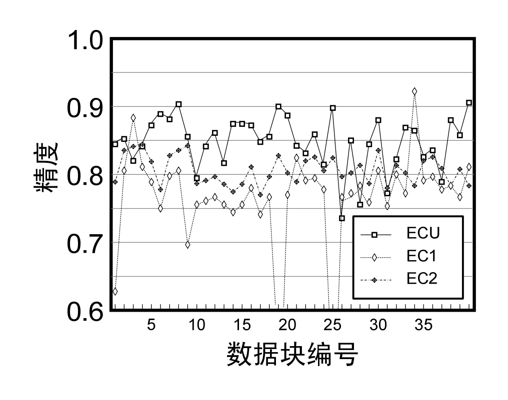
第二步：**决策传播** 当各个基之间的相似性被计算出来以后，接下来的问题就是如何在这些基之间传播相似性了。其目的是要把所有聚类器对应的基（Cluster ID）映射到一个真正的类别标签（Class Label）上。对于一个标签未知的聚类器，我们的基本思想是先把所有的分类器组合起来推导该聚类器的标签，然后利用其他的聚类器来修正该结果。这是因为当分类器很少的时候，仅仅依靠分类器无法推导精确的标签，还需要利用到聚类器之间的结构相似性进行修正。

第三步：**决策协商** 当数据流中潜在模式发生连续变化的时候，各个模型在决策中的作

[3] 杰卡德距离用两个集合中不同元素占所有元素的比例来衡量两个集合的区分度

用将随时间改变。为此，根据数据流中模式的“最近最相似”原则，对各个模型依据其和最近一个模型的相似性进行加权。最后的加权平均将用来进行最后的决策。

通过以上三步，我们可以以较低的代价构筑多通道网络流上的决策（结果）融合模型。如图 3 中所示，在 UCI 恶意网站检测（左图）和 KDDCUP’99 入侵检测（右图）数据集上的测试结果表明，融合多个通道上的信息进行分类的分类模型（ECU），比以往任何单一通道上的分类模型（EC1 和 EC2）精度都要高，注意此处的精度是此被正确分类的数据占全部数据的比例，比例范围是 0 到 1 之间。



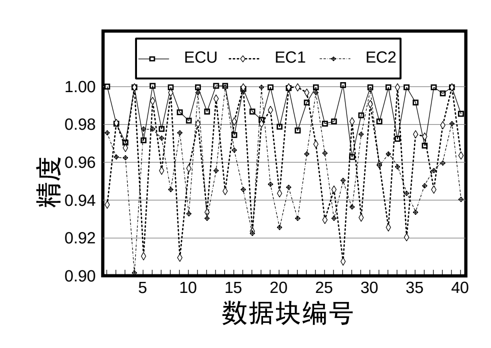


图 3. 在 UCI 恶意网站检测[6]和入侵检测数据集的前 40 个数据块上的对比结果

3.2 基于 R-树的分类模型索引

决策树是一种分类模型。它构造简单，执行速度很快，很适于对网络流进行分类。在实现多通道数据流过滤时，通常如果只用一个决策树在分类精度上难以满足实际应用的需求。集成分类器可以很好地解决这个问题。所谓集成分类器，就是使用多个分类器来对数据进行分类，再对每个分类器得出的结果进行综合得出最终的分类结果。实验显示使用集成分类器来进行多通道数据流过滤可以达到令人满意的精度，但是分类的时间开销会随着分类器数目的增加而线性增加（如图 5 实验结果所示）。在高速网络流环境中，这是不可接受的。通过对集成分类器的决策树利用基于 R-树的索引结构进行索引，可以大大降低分类的时间开销，使集成分类器在高速网络数据流环境中变得可行。

下面介绍 R-树 [7,8,9,10]。R-树是格特曼（A. Guttman）在 1984 年提出的。R-树是经典的索引结构 B-树的多维扩展。R-树和 B-树一样，是一种高度平衡的多路搜索树，是一种外存索引结构。不过，R-树的思想很容易推广到内存索引的情况。R-树的一个结点是若干个索引记录的数组，对于叶结点，它的索引记录具有如下形式：(***I***, ***tuple-ID***)。其中 ***I*** 是一个 ***n*** 维矩形，即一个空间目标的 ***n*** 维矩形表示，***Tuple-ID*** 是一个空间目标的编号。对于非叶结点，它的索引记录具有如下的形式：(***I***, ***child-pointer***)。其中 ***I*** 是一个 ***n*** 维矩形，***child-pointer*** 是一个指向下一级子结点的指针，***I*** 是覆盖了 ***child-pointer*** 所指的结点中所有矩形的最小矩形。

R-树检索过程要解决的问题很简单，就是给定一个 ***n*** 维矩形，判定索引中哪些矩形覆盖了它。R-树的检索过程是从树根往下遍历，但不同的是，由于 R-树每个结点中各个索引记录中的矩形可能重叠，因此需要顺序检查结点中的每个索引记录，这可能导致遍历当前结点的多个子树。所以最坏情况下，R-树的检索可能要访问树中所有的结点。也就是说最坏的情况下，需要将数据样本一个个地与分类规则进行比较。但平均情况下，只需访问 R-树少数几个结点就能完成检索。

由于 R-树是一种针对空间位置的索引，所以只能用来索引连续属性。并且由于高维空间

数据的稀疏性，R-树在处理超高维的数据时存在一些困难。我们的主要工作，在于在 R-树之上增加了一层哈希（Hash）结构，用来对数据中的离散属性进行索引，通过哈希来找到进入 R-树的合适的入口结点，其数据结构如图 4 所示：

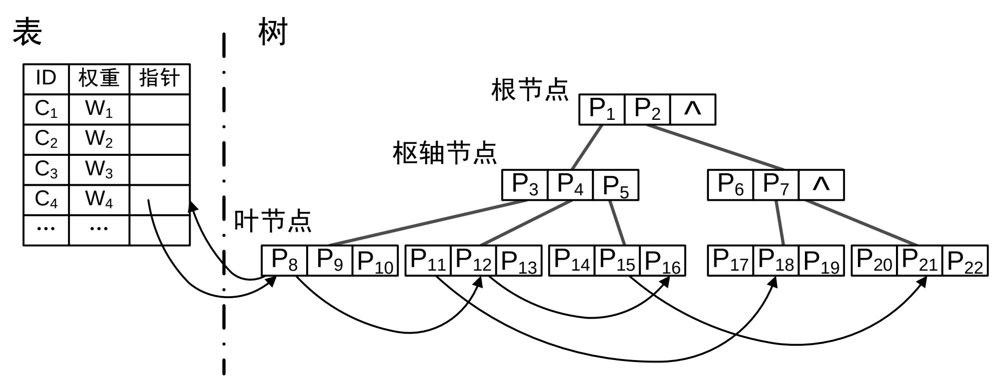


图 4. 基于 R-树的分类模型索引

我们分别在人工数据集（图 5-I）、UCI 恶意网页地址检测数据集（图 5-II）、UCI 垃圾邮件检测数据集（图 5-III）和 KDDCUP'99 入侵检测数据集（图 5-IV）上进行了测试，结果表明使用基于 R-树的索引结构比不使用索引结构时时间开销有显著下降；而且，进行索引后分类的时间开销对集成分类器中分类模型的增加变得不敏感。

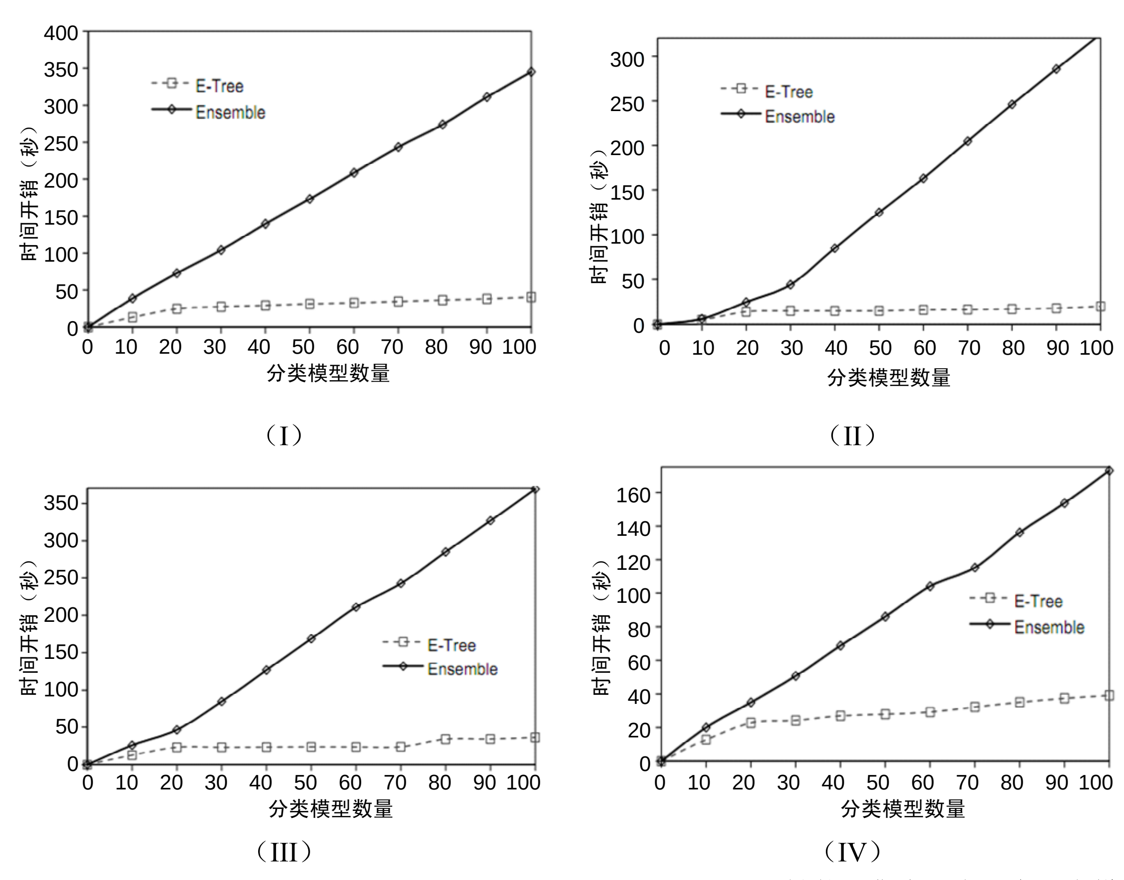


图 5. 进行索引与不进行索引的分类模型在分类速度上的比较（横轴是集成分类器中分类模型的个数，纵轴是对一个数据进行分类的时间开销）

3.3 多通道网络流分类过滤实验平台：F9

真实网络环境具有复杂性与不可预知性，所以很多在理论上很好的算法与模型在真实网

络环境中的性能并不如预期的好，并且由于网络流的复杂性，使得人造实验数据难以模拟真实数据。F9 系统的设计目标就是从真实的网络环境中获取网络数据流，对模型和算法的性能进行检测，同时也可作为真实数据流过滤系统的原型系统。

F9 系统实现了以下功能：

1. **数据流分析**：此功能是利用分析引擎对高速数据流网关中数据流进行有效的协议还原和流拼接，将还原的数据流用分析引擎的匹配算法进行高效的规则匹配分析，检测出满足匹配规则的数据流；
2. **规则库管理**：系统管理员通过配置管理界面动态增删过滤规则，特征提取分析模块将动态生成的规则数据传递给分析引擎，分析引擎动态通过调整内存数据结构使动态规则数据生效；
3. **网络连接阻断**：通过黑白名单的动态设置，利用系统的串联接入方式，实现对特定数据流的实时阻断；
4. **协议还原分析**：通过对网关中数据流进行有效的协议还原，按照协议类别进行有效的内容提取。

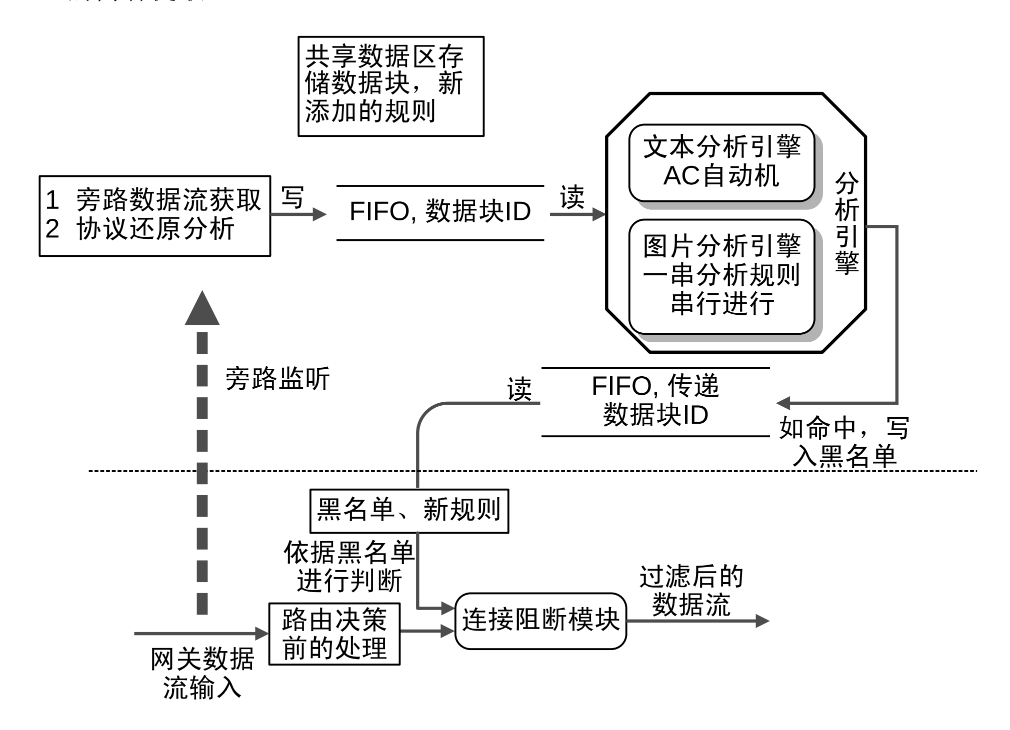


图 6. F9 系统流程图

F9 通过采用串联过滤控制和旁路侦听结合的设计，实现对多媒体流的实时高效的监管控制。系统主要包括：连接阻断模块、协议还原分析模块、数据分析模块、规则发现模块、规则库管理模块，其中：连接阻断模块根据管理的黑名单，对数据流进行有效的地址分析并实施必要拦截阻断；协议还原分析模块利用旁路监听技术将网卡中的媒体流数据导入到数据分析服务器中，然后对导入的原始数据流进行有效的协议分析，识别出数据流所承载的协议，并按照数据流所属协议的规则进行有效的内容抽取，分析其承载的数据内容，如图片、文本、网页地址等；数据分析模块根据协议还原分析模块导入的数据信息类别，分别调用分析引擎内相对独立的分析进程，判断出导入的数据信息是否是敏感数据，如果是敏感数据，则记录其地址信息并导入到连接阻断模块使用的黑名单列表中；规则发现模块利用聚类技术将内容分析服务器中获取的数据流（包括文本和图片）进行快速的抽取和聚集，将聚类进程产生的

结果以页面的形式展现并根据人工选择性标示写入规则库数据集中；规则库管理模块主要是通过界面完成规则库的增删改查操作以及一些辅助的权限控制操作。F9 系统的工作流程如图 6 所示

4　总结

面向过滤的多通道网络流分类技术是网络内容安全处理的核心技术，利用多通道信息对数据流进行过滤在精度和抗破解性上较传统方法有巨大的优势。近年来，我们在多通道网络流分类方面的研究取得了诸多进展，包括开发了准确的流分类模型；构造了高效的模型索引结构；以及开发了面向应用的 F9 多通道网络流过滤平台。这些工作相辅相成，构成了一个有机的整体，为以后在多通道网络流过滤方面的深入研究奠定了基础。

参考文献：

[1] Domingos, P., Hulten, G. (2000) Mining high-speed data streams. Proceedings KDD 2000, ACM Press, New York, NY, USA, pp. 71–80.

[2] Hulten, G.,Spencer, L.,Domingos, P. (2001) Mining time-changing data streams. Proceedings KDD 2001, ACM Press, New York, NY, pp. 97–106.

[3] Gama, J., Fernandes, R., & Rocha, R. (2006) Decision trees for mining data streams. Intelligent Data Analysis 10 23-45.

[4] A. and Babcock, B. and Babu, S. and Cieslewicz, J. and Datar, M. and Ito, K. and Motwani, R and Srivastava, U.and Widom, J. (2004) *STREAM: The Stanford Data Stream Management System.* Technical Report. Stanford InfoLab.

[5] A. Arasu, S. Babu, and J. Widom. The cql continuous query language: Semantic foundations and query execution. Technical report, Stanford University Database Group, Oct. 2003.

[6] UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/KDD+Cup+1999+Data, http://archive.ics.uci.edu/ml/datasets/URL+Reputation

[7] Antonin Guttman: R-Trees: A Dynamic Index Structure for Spatial Searching, Proc. 1984 ACM SIGMOD International Conference on Management of Data, pp. 47–57. ISBN 0-89791-128-8

[8] Yannis Manolopoulos, Alexandros Nanopoulos, Apostolos N. Papadopoulos, Yannis Theodoridis: R-Trees: Theory and Applications, Springer, 2005. ISBN 1-85233-977-2

[9] N. Beckmann, H.-P. Kriegel, R. Schneider, B. Seeger: The R*-Tree: An Efficient and Robust Access Method for Points and Rectangles. SIGMOD Conference 1990: 322-331

[10] Scott T. Leutenegger, Jeffrey M. Edgington and Mario A. Lopez: STR: A Simple and Efficient Algorithm for R-Tree Packing

[11] Dr. Thomas Porter (2005-01-11). "The Perils of Deep Packet Inspection". Security Focus. Retrieved 2008-03-02.

[12] S. Wu and U. Manber, “A fast algorithm for multi-pattern searching”, Dept. of Computer Science, University of Arizona, Tucson, AZ, TR-94-17, 1994

[13] R. K. Srihari, W. Li, C. Niu and T. Cornell,"InfoXtract: A Customizable Intermediate Level Information Extraction Engine",Journal of Natural Language Engineering, Cambridge U. Press , 14(1), 2008, pp.33-69

作者简介：

王　鹏：　中国科学院计算技术研究所信息安全研究中心博士生
wangpeng@software.ict.ac.cn

李　军：　中国科学院计算技术研究所信息安全研究中心博士生

张　鹏：　中国科学院计算技术研究所信息安全研究中心博士生助理研究员